DOCUMENT RESUME

ED 428 122 TM 029 593

AUTHOR Sireci, Stephen G.; Wiley, Andrew; Keller, Lisa A.
TITLE An Empirical Evaluation of Selected Multiple-Choice Item Writing Guidelines.
PUB DATE 1998-10-28
NOTE 19p.; Paper presented at the Annual Meeting of the Northeastern Educational Research Association (Ellenville, NY, October 28-30, 1998).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Certified Public Accountants; Licensing Examinations (Professions); *Multiple Choice Tests; *Test Construction; Test Items
IDENTIFIERS Large Scale Assessment

ABSTRACT

Seven specific guidelines included in the taxonomy proposed by T. Haladyna and S. Downing (1998) for writing multiple-choice test items were evaluated. These specific guidelines are: (1) avoid the complex multiple-choice, K-type format; (2) state the stem in question format; (3) word the stem positively; (4) avoid the phrase "all of the above"; (5) avoid the phrase "none of the above"; (6) avoid specific determiners such as "always" or "never"; and (7) keep the length of options fairly consistent. These guidelines were evaluated by comparing statistical indices of item quality across items that do and do not violate one or more of these guidelines. The items and their statistics were taken from a recently administered, high-stakes, large-scale licensure examination, the Uniform Certified Public Accountant Examination. Only 1 of the 285 items evaluated violated the guidelines to avoid the phrases, "none of the above" and "all of the above"; and the determiners, "always" and "never." The only guideline supported by the data was avoiding the K-type item, since K-type items on this test tended to be more difficult and to have lower discrimination statistics. Results do not support the "state the stem in the question" guideline. Relatively few items violated more than one guideline. (Contains 8 tables and 10 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 428 122

Running Head: Item Writing Guidelines

An Empirical Evaluation of Selected Multiple-Choice Item Writing Guidelines

Stephen G. Sireci
University of Massachusetts–Amherst

Andrew Wiley
American Institute of Certified Public Accountants

Lisa A. Keller
University of Massachusetts–Amherst

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY
StephenSireci
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM029593

Paper presented at the annual conference of the Northeastern Educational Research Association, October 28, 1998, Ellenville, NY

BEST COPY AVAILABLE

Abstract

Numerous guidelines for developing quality multiple-choice test questions appear in the literature. Many of these guidelines, such as those pertaining to use of correct and consistent grammar and appropriateness of the correct answer, are straightforward and their importance for facilitating test validity is unequivocal. Others, such as "word the stem positively" and "state the stem in question format" are less intuitive and their utility is more questionable. This study examined the usefulness of seven popular item writing guidelines by comparing the item difficulty (p-value) and discrimination (biserial) statistics associated with items that violate and do not violate one or more of these guidelines. The items evaluated were from a large-scale testing program comprising 285 multiple-choice items. Several items were identified that violated one or more guidelines. The most frequent violation was using the incomplete stem format rather than the question format. The second most frequent violation was use of the complex (K-type) format. No substantive differences in item difficulty or discrimination were found between the incomplete stem and question formats. However, some evidence was found that the K-type items were slightly more difficult and less discriminating than other items. Few items violated more than one guideline, and it was noted that items that violated more than one guideline tended to be of poorer quality. Future directions for research in this area are discussed.

Introduction

Numerous guidelines for developing quality multiple-choice test items have been published in measurement journals (e.g., Adams, 1992; Haladyna & Downing, 1989a) and textbooks (e.g., Gallagher, 1998; Haladyna, 1994; Roid & Haladyna, 1982; Linn & Gronlund, 1990; Osterlind, 1989). Many of these guidelines, such as those pertaining to use of proper and consistent grammar and appropriateness of the correct answer, are straightforward and their importance for facilitating item quality is unequivocal. Others, such as "word the stem positively" and "state the stem in question format" (Haladyna & Downing, 1989a, pp. 40-41) are less intuitive and their utility is more questionable. Haladyna and Downing (1989b) reviewed the empirical and theoretical literature on the 43 item writing guidelines identified by Haladyna and Downing (1989a) and found that many of the guidelines were unsupported by empirical research. They suggested revisions to other guidelines based on results from empirical analysis of item response data.

Item writing guidelines are taken seriously by testing organizations. Most large-scale test developers include numerous guidelines in the item construction manuals from which their item writers operate. For example, the American Institute of Certified Public Accountants (AICPA) provides its item writers with a 70-page booklet to help them develop quality items (AICPA Board of Examiners, 1995). This booklet includes 23 guidelines for writing multiple-choice items. Although multiple-choice item writing guidelines are credited with promoting high quality items, it is possible that some of these guidelines are not at all useful. As Haladyna and Downing (1989b) concluded: "few item writing rules have received adequate study" and "certain rules appear ... in need of significant new research" (p. 72).

This study examined seven specific item writing guidelines that were included in Haladyna and Downing's (1989a) taxonomy, but are commonly violated by many item writers. The seven specific item writing guidelines evaluated were:

1) avoid the complex (K-type) multiple-choice format
2) state the stem in question format
3) word the stem positively, avoid negative phrasing
4) avoid the phrase "all of the above"
5) avoid the phrase "none of the above"
6) avoid specific determiners such as "always" or "never"
7) keep the length of options fairly consistent.

These guidelines were evaluated by comparing statistical indices of item quality (i.e., item difficulty and item discrimination) across items that do and do not violate one or more of these guidelines. The items, and their statistics, were taken from a recently administered version of a high-stakes, large scale licensure examination.

Method

Data

The data came from the November 1995 administration of the Uniform Certified Public Accountant Examination (CPA Exam). The CPA Exam is a high-stakes test that professional accountants must pass in order to become licensed as a Certified Public Accountant in the United States. The CPA Exam comprised four sections: *Accounting and Reporting* (ARE), *Auditing* (AUD), *Financial Accounting and Reporting* (FARE), and *Law and Professional Responsibilities* (LPR). All sections contain multiple-choice items and other selected-response item types such as multiple true/false items. Three of the sections (all except ARE) also contain constructed-response items. Only data from the multiple-choice items were analyzed in this study. The multiple-choice items make up 50% of the total score on the AUD section and 60% of the total scores on the other sections. There were 285 multiple-choice items across the four test sections. The number of items on each section ranged from 60 to 90. About 51,000 candidates took each section of the test. Table 1 provides the number of items, sample sizes, and some descriptive statistics for each test section. The mean item difficulties among the four sections ranged from .51 to .63. The mean discrimination statistics were more homogeneous, ranging from .36 to .41.

[Insert Table 1 Here]

Procedure

A content analysis was performed on the 285 multiple-choice items to identify items that violated one or more of the item writing guidelines. Seven dichotomous dummy variables were created to indicate whether an item violated any of the seven guidelines. These dichotomous data served as grouping variables to compare statistical indices of item difficulty and discrimination across items that violated and did not violate one or more of the guidelines. All candidates who responded to an item were included in calculating the item statistics; thus, the statistics were derived from sample sizes of about 51,000. The difficulty index was the unadjusted "p-value," which is the proportion of candidates who answered the item correctly divided by the total number of candidates who responded to the item. The discrimination index was the biserial correlation between the dichotomous item score (i.e., 0=incorrect, 1=correct) and the total score on the remainder of the multiple-choice items. Descriptive statistics, t-tests, and one-way analyses of variance (ANOVAs) were performed separately for each section. Multiple comparison procedures were used where appropriate to compare the statistics for items that violated more than one guideline with those that violated fewer guidelines.

Results

Table 2 presents a detailed summary of the numbers of items that violated each guideline per section. Several findings are notable. First, none of the 285 items violated the "avoid using

the phrase all of the above" or "avoid using the phrase none of the above" guidelines. Obviously, the AICPA did not allow those types of items on the test. Second, only one item violated the "avoid specific determiners such as 'always' or 'never'" guideline. A third interesting finding is that the most frequently violated guideline was "state the stem in question format." In fact, two-thirds of the AUD items used the incomplete stem format. The next most frequently violated guideline was use of the complex multiple-choice (K-type) format. The largest number of items violating this guideline were found in the ARE section. The next most frequently violated guidelines were "keep the length of options fairly consistent," and "word the stem positively, avoid negative phrasing."

[Insert Table 2 Here]

Table 3 summarizes the number of item writing guideline violations by subtest. The AUD section had the largest number of item writing guideline violations due to the frequent use of the incomplete stem format. Very few items violated more than one guideline (only 18 across all four sections), and in all such cases the items violated only two guidelines. ARE had the largest number of double violations (9), which represented 12% of the items.

[Insert Table 3 Here]

The numbers of items per section that violated specific guidelines precluded the ability to evaluate all of the guidelines statistically. However, several analyses were conducted where sample sizes permitted. The analyses focusing on the "state the stem in question format" guideline are summarized in Table 4 (item difficulty comparison) and Table 5 (item discrimination comparison). The only notable finding was that the incomplete stem ARE items were statistically significantly easier than the question format ARE items. There were no differences in difficulty between items that did and did not violate this guideline for the other test sections. The discrimination comparisons (Table 5) discovered no differences among the discrimination statistics for incomplete stem and question format items. These findings do not support the utility of this guideline.

[Insert Table 4 Here]

[Insert Table 5 Here]

The results of the evaluation of the "K-type" guideline are presented in Table 6 (item difficulty) and Table 7 (item discrimination). The numbers of items violating this guideline within a subtest were too small to allow for statistical comparison; thus, only descriptive statistics are presented. It is interesting to note that the K-type items are more difficult for three of the four sections. The largest difference observed was for AUD, although there were only 3 K-type items in this section. The mean p-value for these three items was .38, which was .25 lower than the average of the other 87 items. Given this small sample size, not much can be generalized from this result. However, the 11 K-type ARE items had an average p-value that was .12 lower than

the other 64 ARE items. These findings are consistent with the literature, which states that K-type items tend to be more difficult. The p-value differences for the FARE and LPR tests were minor, with the 6 LPR K-type items exhibiting a higher (i.e., easier) mean p-value.

[Insert Table 6 Here]

Similar results for the K-type guideline were observed with respect to item discrimination. Three average discrimination statistics for the K-type items were lower than those of the other items for three of the four sections. This finding, taken together with the difficulty differences noted above, supports previous findings that claim the K-type format may be confusing for test takers, making them more difficult and less discriminating.

[Insert Table 7 Here]

Descriptive statistics are not presented for the items that violated the other guidelines because, within any section, they are so few in number. However, additional analyses were carried out to compare items that violated two guidelines with those that did not violate a guideline or violated only one guideline. These analyses are important because there may be a cumulative effect when an item violates more than one guideline.

As illustrated in Table 3, there were 18 items that violated two guidelines. Sixteen of these items involved an incomplete stem item coupled with another violation. For each section, item difficulty and discrimination statistics were compared across items that had zero, one, or two violations. Items were categorized into one of these three groups and one-way ANOVAs were computed using p-values and biserials as the dependent variables. Four planned comparisons were conducted for each analysis. All comparisons used a Bonferroni-adjusted alpha of .05. The first three comparisons reflected the three possible pairwise comparisons. The fourth comparison compared the item statistics for those items violating two guidelines with those items violating zero or one guideline. Although the sample sizes were different across the three groups, all ANOVAs met the homogeneity of variance assumption as tested by the Levene statistic.

The multiple-group comparisons are summarized in Table 8. This table provides the item difficulty and discrimination means and standard deviations for all groups of items. It also details the types of "double violation" items in each section. The descriptive statistics illustrate two findings that support the hypothesis that items violating more than one guideline have poorer statistics. First, the mean item discrimination for the 9 "double violation" ARE items (.25) is noticeably lower than the means for the items that violated one or fewer guidelines (.42 and .36, respectively). Second, the double violation items had lower average item difficulties for the three sections that had at least two double violation items. However, only the first finding was supported by the statistical analyses. The average discrimination of the double violation ARE items was statistically significantly lower ($p < .05$) than the average discrimination of the ARE items that violated only one guideline. The only other statistically significant finding was that the ARE items that violated only one guideline were statistically significantly easier than the ARE

items that did not violate any guidelines, or that violated two guidelines. No other comparisons, and none of the planned contrasts, were statistically significant.

[Insert Table 8 here]

Discussion

This study provided several interesting findings regarding the validity of the item writing guidelines studied. First, it is interesting that three of the guidelines were strictly adhered to by the AICPA: avoid the phrase "all of the above," avoid the phrase "none of the above," and avoid specific determiners such as "always" or "never." Only one of the 285 items violated one of these guidelines. Second, the only guideline supported by the data was "avoid the complex (K-type) multiple-choice format." The K-type items on this test tended to be more difficult and have lower discrimination statistics. A third interesting finding was that the incomplete stem items tended to have difficulty and discrimination statistics that were as good as or better than the question format items. Thus, the present study does not support the "state the stem in question format" guideline.

Perhaps the most interesting finding of the study was the relatively small numbers of items that violated more than one guideline. No item violated more than two guidelines and only 18 of the 285 items (6%) violated two. The majority of these double violation items (16 of 18) were incomplete stem items. Preliminary evidence was provided that when an item violates more than one guideline, item quality may diminish. Thus, although the incomplete stem format may not be problematic, it may cause confusion when coupled with the K-type format, or when negative phrasing is used in the stem or response options. However, the small number of double violation items within a test obviated thorough statistical analysis of this condition.

Haladyna and Downing (1989b) reviewed 96 theoretical and empirical studies that appraised item writing guidelines. Our results are consistent with their finding that 8 out of 9 studies concluded the K-type format was more difficult. However, our results do not support their revised guideline "use the question format, avoid the completion format." This guideline deserves further study, preferably using an experimental research design.

Although the results of this study provide important data regarding the predominance of items that violate the guidelines and the consequences associated with such violations, there are several limitations to be noted. First, the design of the study is non-experimental. Future research should explore constructing parallel versions of items that do and do not violate one or more guidelines and administering these items to test takers. Doing so would allow for an experimental design that could better control for content and context differences among the items. Another limitation of this study is that the results may not be generalizable beyond the CPA Exam. Great care is taken in developing the CPA Exam. The test development window is about 18 months and each item is scrutinized several times before appearing on an exam. Thus, the items evaluated in this study that violated a guideline may not be "typical" of items that violate a guideline and receive less scrutiny. Items from other exams should be analyzed to see if the similar results are

obtained. A further limitation is that only one CPA Exam test form was evaluated. We hope to extend this research by evaluating other test forms.

The use of item difficulty and discrimination statistics to evaluate item quality also has limitations. Items that violate item writing guidelines may undermine test validity in ways that do not show up in item statistics. For example, some items may facilitate test anxiety. Others may take longer to answer. If these items appear earlier on the test, the smarter test takers may still answer them correctly, but their performance on later items may be affected. One potential way to evaluate this problem would be to administer items on a computer and compare response times across items that do and do not violate these guidelines. Another potential method would be to use "think-aloud" protocols or interviews of test takers who respond to different item formats.

Gross (1994) addressed the issue of logical versus empirical guidelines for writing multiple-choice items. He argued that some item writing guidelines, such as "avoid the phrase all of the above," and "avoid the phrase none of the above," are defensible logically and are not in need of empirical support. As he stated, "any stem or option format that by design diminishes an item's ability to distinguish between candidates with full versus misinformation should not be used" (p. 125). This advice appears to be adhered to by the AICPA. Future research should focus on gathering empirical support for those guidelines that are not so logically defensible, such as "state the stem in question format."

References

Adams, R. (1992. February). Multiple-choice item writing: Art and Science. The Bar Examiner. 5-14.

American Institute of Certified Public Accountants Board of Examiners (1995). Uniform CPA Examination preparation Guide. New York, NY: American Institute of Certified Public Accountants.

Gallagher, J. D. (1998). Classroom assessment for teachers. Upper Saddle River, NJ: Merrill.

Gross, L. J. (1994). Logical versus empirical guidelines for writing test items. Evaluation & The Health Professions, 17, 123-126.

Haladyna, T. M. (1994). Developing and validating multiple-choice test items. Hillsdale: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1989a). A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 37-50.

Haladyna, T. M., & Downing, S. M. (1989b). Validity of a taxonomy of multiple-choice item-writing rules. Applied Measurement in Education, 2, 51-78.

Linn, R. L., & Gronlund, N. E. (1995). Measurement and assessment in teaching (7th edition). Englewood Cliffs, NJ: Merrill.

Osterlind, S. J. (1989). Constructing test items. Hingham, MA: Kluwer.

Roid, G. H., & Haladyna, T. M. (1982). A technology for test-item writing. Orlando, FL: Academic Press.

Table 1

Descriptive Statistics For CPA Exam Subtests

	ARE	AUD	FARE	LPR
Number of Items	75	90	60	60
Number of Candidates	51,548	51,998	50,825	51,915
Mean Difficulty (SD)	.51 (.22)	.62 (.14)	.53 (.18)	.63 (.18)
Mean Discrimination (SD)	.37 (.21)	.41 (.16)	.39 (.16)	.36 (.16)
KR-20 Reliability	.85	.89	.82	.77

Table 2

Numbers of Item Writing Guideline Violations by Subtest

Guideline Violated	ARE (75 Items)	AUD (90 Items)	FARE (60 Items)	LPR (60 Items)	**Total (285 Items)**
Question format	32	60	11	18	**121**
Avoid "K-type"	11	3	2	6	**22**
Keep length consistent	3	2	2	2	**9**
Avoid negative phrasing	0	6	2	1	**9**
Avoid "Always," "Never"	0	0	0	1	**1**
Avoid "All of the Above"	0	0	0	0	**0**
Avoid "None of the Above"	0	0	0	0	**0**

Table 3

Tabulations and Percentages of Single and Double Item Writing Guideline Violations

	ARE (75 Items)	AUD (90 Items)	FARE (60 Items)	LPR (60 Items)	**TOTAL[a] (285 Items)**
Total # "Problem" Items	37	65	15	27	**144**
% Items Violating At Least One Guideline	49%	72%	25%	45%	**51%**
# Items Violating Two Guidelines	9	6	2	1	**18**
% Items Violating Two Guidelines	12%	7%	3%	2%	**6%**

[a]Numbers represent total items summed across test sections; percentages represent total number of items across all four sections divided by the total number of items (285).

Table 4

Comparison of **Item Difficulty** Differences: Question Format vs. Incomplete Stem

	Question Format			Incomplete Stem					
Subtest	n	Mean	(SD)	n	Mean	(SD)	t	η^2	CI_{95}
ARE	43	.45	(.20)	32	.59	(.21)	-2.95*	.107	{-.236, -.046}
AUD	30	.62	(.16)	60	.62	(.14)	.026	.000	{-.064, .065}
FARE	49	.53	(.19)	11	.53	(.15)	.029	.000	{-.119, .123}
LPR	42	.64	(.17)	18	.59	(.22)	.883	.013	{-.058, .150}

Table 5

Comparison of **Item Discrimination** Differences: Question Format vs. Incomplete Stem

	Question Format			Incomplete Stem					
Subtest	n	Mean	(SD)	n	Mean	(SD)	t	η^2	CI_{95}
ARE	43	.35	(.22)	32	.39	(.20)	-.893	.010	{-.140, .053}
AUD	30	.41	(.15)	60	.41	(.16)	-.232	.001	{-.080, .063}
FARE	49	.38	(.16)	11	.40	(.14)	-.731	.009	{-.132, .080}
LPR	42	.35	(.17)	18	.38	(.15)	-.494	.004	{-.125, .058}

Table 6

Comparison of **Item Difficulty** Differences: "K-Type" Items

	NOT K-Type Format		K-Type Format		
Subtest	n	Mean (SD)	n	Mean (SD)	Mean Difference
ARE	64	.52 (.21)	11	.40 (.21)	.12
AUD	87	.63 (.14)	3	.38 (.08)	.25
FARE	58	.54 (.18)	2	.49 (.15)	.05
LPR	54	.62 (.18)	6	.68 (.23)	-.06

Table 7

Comparison of **Item Discrimination** Differences: "K-Type" Items

	NOT K-Type Format		K-Type Format		
Subtest	n	Mean (SD)	n	Mean (SD)	Mean Difference
ARE	64	.38 (.21)	11	.25 (.18)	.13
AUD	87	.42 (.16)	3	.28 (.16)	.14
FARE	58	.38 (.16)	2	.41 (.16)	-.03
LPR	54	.37 (.17)	6	.30 (.11)	.07

Table 8

Comparisons of Items Violating Zero, One, and Two Guidelines

Subset	n[a]	Difficulty	Discrimination	Description of "Double Violation" Items
		Mean (SD)	Mean (SD)	
ARE				
Zero violations	38	.45 (.20)	.36 (.23)	
One violation	18	.61 (.21)	.42 (.15)	
Two violations	9	.41 (.20)	.25 (.23)	6 Inc. Stem/K-type, 3 Inc. Stem/"Improper Length"
AUD				
Zero violations	25	.62 (.16)	.40 (.17)	
One violation	59	.62 (.14)	.42 (.16)	
Two violations	6	.57 (.17)	.43 (.19)	2 Inc. Stem/K-type, 2 Inc. Stem/"Improper Length," 2 Inc. Stem/"Negative Phrasing"
FARE				
Zero violations	45	.53 (.18)	.38 (.17)	
One violation	13	.57 (.18)	.39 (.14)	
Two violations	2	.49 (.15)	.41 (.16)	2 Inc. Stem/K-type
LPR				
Zero violations	33	.62 (.16)	.36 (.18)	
One violation	26	.63 (.22)	.36 (.14)	
Two violations	1	.74	.42	1 Inc. Stem/"Improper Length"

[a]n indicates the number of items in each category.